A PROPOSAL OF FUZZY MULTIDIMENSIONAL ASSOCIATION RULES

Rolly Intan
Department of Informatics Engineering, Petra Christian University, Surabaya
Email: rintan@petra.ac.id

ABSTRACT: Association rules that involve two or more dimensions or predicates can be referred as multidimensional association rules. Rather than searching for frequent itemsets (as is done in mining single-dimensional association rules), in multidimensional association rules, we search for frequent predicate sets. In general, there are two types of multidimensional association rules, namely interdimension association rules and hybrid-dimension association rules. Interdimension association rules are multidimensional association rules with no repeated predicates. This paper introduces a method for generating interdimension association rules. A more meaningful association rules can be provided by generalizing crisp value of attributes to be fuzzy value. To generate the multidimensional association rules implying fuzzy value, this paper introduces an alternative method for mining the rules by searching for the predicate sets.

Keywords: Multidimensional Association rules, Inter- dimension Association Rules, Data Mining, Fuzzy Sets.

INTRODUCTION

Association rule mining finds interesting association or correlation relationship among a large data set of items [1,8,11]. The discovery of interesting association rules can help in decision making process.

Association rule mining that implies a single predicate is referred as a single dimensional or *intradimension association rule* since it contains a single distinct predicate with multiple occurrences (the predicate occurs more than once within the rule). The terminology of *single dimensional or intradimension association rule* is used in multidimensional database by assuming each distinct predicate in the rule as a dimension. For instance, in *market basket analysis*, customers buying habit is analyzed for finding association between different items customers put together in their shopping cart. In *market basket analysis*, it might be discovered a Boolean association rule "laptop $\Rightarrow$ b/w printer" which can also be written as a single dimensional association rule as follows [1]:

Rule-1

buys (*X*, "laptop") $\Rightarrow$ *buys* (*X*, "b/w printer"),

where *buys* is a given predicate and *X* is a variable representing customers who purchased items (e.g. *laptop* and *b/w printer*). In general, *laptop* and *b/w printer* are two different data that are taken from a certain database attribute, called *items*. In general, *Apriori* [1,8] is used an influential algorithm for mining frequent itemsets for generating Boolean (single dimensional) association rules. To provide a more human-based concept, in [6,11] we proposed an alternative algorithm for generating the (single dimensional) association rule by utilizing fuzzy sets in the market basket analysis.

Additional relational information regarding the customers who purchased the items, such as customer age, occupation, credit rating, income and address, may also have a correlation to the purchased items. Considering each database attribute as a predicate, it can therefore be interesting to mine association rules containing *multiple* predicate, such as:

Rule-2

$$age(X,"20") \wedge occupation(X,"student") \Rightarrow buys(X,"laptop"),$$

where there are three predicates, namely *age*, *occupation* and *buys*. Association rules that involve two or more dimensions or predicates can be referred to as *multidimensional association rules*.

To provide a more meaningful association rule, it is necessary to utilize *fuzzy sets* over a given database attribute as discussed in [13]. Formally, given a crisp domain attribute *D*, any arbitrary fuzzy set (say, fuzzy set *A*) is defined by a membership function of the form [2,7]:

$$A: D \rightarrow [0,1]. \qquad (1)$$

A fuzzy set may be represented by a meaningful fuzzy label. For example, "young" is a fuzzy set over *age* that is defined on the interval [0, 100], and "high_inc" is a fuzzy set over *income* on the interval [100, 1000] as arbitrarily given by:

$$young(x) = \begin{cases} 1 & \text{when } x \le 20 \\ (35-x)/15 & \text{when } 20 < x < 35 \\ 0 & \text{when } x \ge 35 \end{cases}$$

$$high\_inc(x) = \begin{cases} 0 & \text{when } x \le 300 \\ (x-300)/300 & \text{when } 300 < x < 600 \\ 1 & \text{when } x \ge 600 \end{cases}$$

Using the previous definition of fuzzy sets on *age* and *income*, an example of *multidimensional association rule* relation among the predicates *age*, *income* and *buys* may then be represented by:

Rule-3

$$age(X,\text{"young"}) \wedge income(X,\text{"high\_inc"}) \Rightarrow buys(X,\text{"car"}),$$

To generate *multidimensional association rules* implying fuzzy value as given by the above example, this paper introduces an alternative method. The method considered as an extended concept of our previous algorithm proposed in [11]. Two important formulas are introduced to calculate *support* and *confidence factor* for every association rule.

The structure of the paper is the following. In Section II, basic definition and formulation of association rules, support and confidence rule are briefly recalled. Section III as a main contribution of this paper is devoted to propose a concept for generating *multidimensional association rules* mining. Section IV demonstrated the concept in an illustrative example. Finally a conclusion is given in Section V.

ASSOCIATION RULES, SUPPORT AND CONFIDENCE

Association rules are kind of patterns representing correlation of attribute-value in a given set of data provided by a process of data mining system. Generally, association rule is a conditional statement (such kind of *if-then rule*). More formally [2], association rules are the form $A \Rightarrow B$, that is, $A_1 \wedge \Lambda \wedge A_m \Rightarrow B_1 \wedge \Lambda \wedge B_n$, where A_i (for $i \in \{1,\ldots,m\}$) and B_j (for $j \in \{1,\ldots,n\}$) are attribute-value pairs. The association rule $A \Rightarrow B$ is interpreted as " database tuples that satisfy the conditions in A are also likely to satisfy the conditions in B." Performance of an association rule is determined by two factors, namely *confidence* and *support* factors. Confidence is a measure of certainty to assess the validity of the rule. Given a set of relevant data tuples (or transactions in a transaction database) the confidence of "$A \Rightarrow B$" is defined by:

$$\text{confidence}(A \Rightarrow B) = \frac{\#tuples(A \text{ and } B)}{\#tuples(A)}, \quad (2)$$

where #*tuples*(A and B) means the number of tuples containing A and B.

For example, a confidence 80% for the Association Rule (Rule-1) means that 80% of all customers who purchased a laptop also bought b/w printer. The support of an association rule refers to the percentage of relevant data tuples (or transactions) for which the pattern of the rule is true. For the association rule "$A \Rightarrow B$" where A and B are the sets of items, support of the rule can be defined by

$$\text{support}(A \Rightarrow B) = \text{support}(A \cup B) = \frac{\#tuples(A \text{ and } B)}{\#tuples(all\_data)}, \quad (3)$$

where #*tuples*(*all_data*) is the number of all tuples in the relevant data tuples (or transactions).

For example, a support 30% for the Association Rule (Rule-1) means that 30% of all customers in the all transactions purchased both laptop and b/w printer. From (3), it can be followed support $(A \Rightarrow B) = \text{support}(B \Rightarrow A)$. Also, (2) can be calculated by

$$\text{confidence}(A \Rightarrow B) = \frac{\text{support}(A \cup B)}{\text{support}(A)}, \quad (4)$$

A data mining system has the potential to generate a huge number of rules in which not all of the rules are interesting. Here, there are several objective measures of rule interestingness. Two of them are measure of rule support and measure of rule confidence. In general, each interestingness measure is associated with a threshold, which may be controlled by the user. For example, rules that do not satisfy a confidence threshold (*minimum confidence*) of, say 50% can be considered uninteresting. Rules below the threshold (*minimum support* as well as *minimum confidence*) likely reflect noise, exceptions, or minority cases and are probably of less value.

MULTIDIMENSIONAL ASSOCIATION RULES

As explained in the previous section association rules that involve two or more dimensions or predicates can be referred to as *multidimensional association rules*. Multidimensional rules with no repeated predicates are called *interdimension association rules* (e.g. Rule-2)[1]. On the other hand, *multidimensional association rules* with repeated predicates, which contain multiple occurrences of some predicates, are called *hybrid-dimension association rules*. The rules may be also considered as combination (hybridization) between *intradimension association rules* and *interdimension association rules*. An example of such a rule is the following, where the predicate *buys* is repeated:

Rule-4:

$$age(X,\text{"20"}) \wedge buys(X,\text{"laptop"}) \Rightarrow buys(X,\text{"b/w printer"}).$$

Here, we may firstly be interested in mining multidimensional association rules with no repeated

predicates or *interdimension association rules*. *Hybrid-dimension association rules* as an extended concept of *multidimensional association rules* will be discussed later in our next paper.

The *interdimension association rules* may be generated from a relational database or data warehouse with multiple attributes by which each attribute is associated with a predicate. A relational database [12] R consists of a set of tuples, where t_i represents the i-th tuple and if there are n domain attributes D, then $t_i = (d_{i1}, d_{i2}, \Lambda\ , d_{in})$. Here, d_{ij} is an atomic value of tuple t_i with the restriction to the domain D_j, where $d_{ij} \in D_j$. Formally, a relational database R is defined as a subset of the set of cross product $D_1 \times D_2 \times \Lambda \times D_n$. Tuple t (with respect to R) is an element of R. In general, R can be shown in Table 1.

Table 1. A Relational Database

Tuples	D_1	D_2	Λ	D_n
t_1	d_{11}	d_{12}	Λ	d_{1n}
t_2	d_{21}	d_{22}	Λ	d_{2n}
M	M	M	O	M
t_r	d_{r1}	d_{r2}	Λ	d_{rn}

To generate the *multidimensional association rules*, we introduce an alternative method for mining the rules by searching for the predicate sets. Conceptually, a multidimensional association rule, $A \Rightarrow B$ consists of A and B as two datasets, called premise and conclusion, respectively.

Formally, A is a dataset consisting of several distinct data, where each data value in A is taken from a distinct domain attribute in D as given by:

$$A = \{a_j \mid a_j \in D_j, \text{ for some } j \in \mathrm{N}_n\},$$

where, $D_A \subseteq D$ is a set of domain attributes in which all data values of A come from.
Similarly,

$$B = \{b_j \mid b_j \in D_j, \text{ for some } j \in \mathrm{N}_n\},$$

where, $D_B \subseteq D$ is a set of domain attributes in which all data values of B come from.
For example, from Rule-2, it can be found that A={20, student}, B={laptop}, D_A={*age*, *occupation*} and D_B={*buys*}.

Considering $A \Rightarrow B$ is an *interdimension association rule*, it can be proved that $|D_A| = |A|$, $|D_B| = |B|$ and $D_A \cap D_B = \varnothing$.

Support of A is then defined by:

$$\text{support}(A) = \frac{|\{t_i \mid d_{ij} = a_j, \forall a_j \in A\}|}{r}, \quad (5)$$

where r is the number of records or tuples (see Table 1).

Alternatively, r in (5) may be changed to $|QD(D_A)|$ by assuming that records or tuples, involved in the process of mining association rules are records in which data values of a certain set of domain attributes, D_A, are not null data. Hence, (5) can be also defined by:

$$\text{support}(A) = \frac{|\{t_i \mid d_{ij} = a_j, \forall a_j \in A\}|}{|QD(D_A)|}, \quad (6)$$

where $QD(D_A)$, simply called *qualified data* of D_A, is defined as a set of record numbers (t_i) in which all data values of domain attributes in D_A are not null data. Formally, $QD(D_A)$ is defined as follows.

$$QD(D_A) = \{t_i \mid t_i(D_j) \neq null, \forall D_j \in D_A\}. \quad (7)$$

Similarly,

$$\text{support}(B) = \frac{|\{t_i \mid d_{ij} = b_j, \forall b_j \in B\}|}{|QD(D_B)|}. \quad (8)$$

As defined in (3), $\text{support}(A \Rightarrow B)$ is given by:

$$\text{support}(A \Rightarrow B) = \text{support}(A \cup B) = \frac{|\{t_i \mid d_{ij} = c_j, \forall c_j \in A \cup B\}|}{|QD(D_A \cup D_B)|} \quad (9)$$

$\text{confidence}(A \Rightarrow B)$ as a measure of certainty to assess the validity of $A \Rightarrow B$ is calculated by

$$\text{confidence}(A \Rightarrow B) = \frac{|\{t_i \mid d_{ij} = c_j, \forall c_j \in A \cup B\}|}{|\{t_i \mid d_{ij} = a_j, \forall a_j \in A\}|} \quad (10)$$

If support(A) is calculated by (5) and denominator of (9) is changed to r, clearly, (10) can be proved having relation as given by (4).

A and B in the previous discussion are datasets in which each element of A and B is an atomic crisp value. To provide a generalized multidimensional association rules, instead of an atomic crisp value, we may consider each element of the datasets to be a dataset of a certain domain attribute. Hence, A and B are sets of set of data values. For example, the rule may be represented by

Rule-5:

$$age(X, \text{"20...29"}) \wedge income(X, \text{"400...900"}) \Rightarrow buys(X, \text{"car, house"}),$$

where A={{20…29}, {400…900}} and B={{car, house}}. Simply, let A be a generalized dataset. Formally, A is given by

$$A = \{A_j \mid A_j \subseteq D_j, \text{ for some } j \in \mathrm{N}_n\}.$$

Corresponding to (6), support of A is then defined by:

$$\text{support}(A) = \frac{|\{t_i \mid d_{ij} \subseteq A_j, \forall A_j \in A\}|}{|QD(D_A)|}. \quad (11)$$

Similar to (9),

$$\text{support}(A \Rightarrow B) = \text{support}(A \cup B) = \frac{|\{t_i \mid d_{ij} \subseteq C_j, \forall C_j \in A \cup B\}|}{|QD(D_A \cup D_B)|} \quad (12)$$

Finally, $\text{confidence}(A \Rightarrow B)$ is defined by

$$\text{confidence}(A \Rightarrow B) = \frac{|\{t_i \mid d_{ij} \subseteq C_j, \forall C_j \in A \cup B\}|}{|\{t_i \mid d_{ij} \subseteq A_j, \forall A_j \in A\}|} \quad (13)$$

To provide a more generalized *multidimensional association rules*, we may consider A and B as sets of fuzzy labels. Here, we can use meaningful fuzzy label to provide a more meaningful association rules. Simply, A and B are called fuzzy datasets. Rule-3 is an example of such rules, where A={*young*, *high_inc*} and B={car}. A fuzzy dataset is a set of fuzzy data consisting of several distinct fuzzy labels, where each fuzzy label is represented by a fuzzy set on a certain domain attribute. Let A be a fuzzy dataset. Formally, A is given by

$$A = \{A_j \mid A_j \in \mathrm{F}(D_j), \text{ for some } j \in \mathrm{N}_n\},$$

where $\mathrm{F}(D_j)$ is a fuzzy power set of D_j, or in other words, A_j is a fuzzy set on D_j. Corresponding to (6), support of A is then defined by:

$$\text{support}(A) = \frac{\sum_{i=1}^{r} \inf_{A_j \in A} \{\mu_{A_j}(d_{ij})\}}{|QD(D_A)|}. \quad (14)$$

Similar to (9),

$$\text{support}(A \Rightarrow B) = \text{support}(A \cup B) = \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \{\mu_{C_j}(d_{ij})\}}{|QD(D_A \cup D_B)|} \quad (15)$$

Finally, $\text{confidence}(A \Rightarrow B)$ is defined by

$$\text{confidence}(A \Rightarrow B) = \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \{\mu_{C_j}(d_{ij})\}}{\sum_{i=1}^{r} \inf_{A_j \in A} \{\mu_{A_j}(d_{ij})\}} \quad (16)$$

Similarly, If denominators of (14) and (15) are changed to r (the number of tuples), (16) can be proved also having relation as given by (4). Here, we may consider and prove that (15) and (16) are generalization of (12) and (13), respectively. On the other hand, (12) and (13) are generalization of (9) and (10).

ILLUSTRATIVE EXAMPLE

An illustrative example is given to understand well the concept of the proposed method and how to calculate support and confidence of the multidimensional association rule mining is performed. The process is started from a given transactional database as shown in Table 2.

Table 2. Transactional Database

Tuples	Ages	Income	Occupation	Buys
t_1	20	150	Student	Laptop
t_2	25	650	Consultant	Car
t_3	22	250	Student	Laptop
t_4	27	550	Lecturer	House
t_5	30	550	Lecturer	Car
t_6	45	700	Director	House
t_7	40	650	Manager	House
t_8	50	650	Professor	Car
t_9	20	200	Student	Printer
t_{10}	55	750	Director	House
t_{11}	*null*	550	Consultant	Car

Based on Table 2, support and confidence of Rule-2 are calculated using (9) and (10), respectively. Related to the conceptual form of the rule $A \Rightarrow B$, it can be followed that A={20, student} and B={laptop}.

$$\text{support (Rule-2)} = \frac{|\{t_1\}|}{|\{t_1,...,t_{10}\}|} = 0.1,$$

where $QD(D_A \cup D_B) = \{t_1,...,t_{10}\}$. t_{11} is not included in $QD(D_A \cup D_B)$, because it has a null value in *Ages*. Confidence of Rule-2 is given by

$$\text{confidence(Rule - 2)} = \frac{|\{t_1\}|}{|\{t_1, t_9\}|} = 0.5.$$

Support and confidence of Rule-5 are calculated using (12) and (13) as follows.

$$\text{support (Rule-5)} = \frac{|\{t_2, t_4\}|}{|\{t_1,...,t_{10}\}|} = 0.2,$$

$$\text{confidence(Rule - 5)} = \frac{|\{t_2, t_4\}|}{|\{t_2, t_4\}|} = 1.$$

Rule-3 is a fuzzy rule, where A={*young*, *high_inc*} and B={car}. *Young* (*yg*) and *high_inc* (*hi*)

are two fuzzy labels represented by two fuzzy sets as given in Section I. Support of Rule-3 can be calculated by (15) as shown in the following table.

Table 3. Calculation of Fuzzy Values

Tuples	$\mu_{yg}(ages)$ α	$\mu_{hi}(income)$ β	$\mu_{car}(buys)$ γ	Min(α,β,γ)
t_1	1	0	0	0
t_2	0.66	1	1	0.66
t_3	0.87	0	0	0
t_4	0.53	0.83	0	0
t_5	0.33	0.83	1	0.33
t_6	0	1	0	0
t_7	0	1	0	0
t_8	0	1	1	0
t_9	1	0	0	0
t_{10}	0	1	0	0
t_{11}	*null*	0.83	1	0
Σ	4.4	7.5	3	1

Therefore,

$$\text{support(Rule - 3)} = \frac{0.66 + 0.33}{|\{t_1,\ldots,t_{10}\}|} = 0.1.$$

On the other hand, confidence of Rule-3 is given by:

$$\text{confidence(Rule - 3)} = \frac{0.66 + 0.33}{0.66 + 0.53 + 0.33} = 0.65.$$

Other interesting rules can be generated from Table 2 as follows.

Rule-6

$income(X, \text{"high\_inc"}) \Rightarrow occupation(X, \text{"consultant , director"})$
(support = 0.35, confidence = 0.51).
On the other hand,

Rule-7

$occupation(X, \text{"consultant, director"}) \Rightarrow income(X, \text{"high\_inc"})$
(support = 0.35, confidence = 0.96).

It can be concluded that the rule "if someone is a director or consultant, he will receive high income", is more sure than the rule "if someone received high income, he must be a director or consultant". The conclusion is reasonable, since from Table 2, clearly it is not only directors and consultants who received high income, but also manager and professor.

The proposed concept and method had been applied in our research to analyze medical track record patients in order to know distribution of a certain disease [10]. In the application, the data of medical track record patients consist of *Date*,*Allergy*,*Pre-diagnosis*, *Post-diagnosis*, *Disease*,*Address*,*Occupation*,*Ages*,*Sex*,*Religion*,*Blood Type*,etc. Let, *D* is a fuzzy label on diseases, *Al* is a fuzzy set on allergy, and *Ad* is a fuzzy set on address that can be meant an area or location of residence. Result of fuzzy association rule can be given by the following pattern:
$address(X, \text{"}Ad\text{"})$ and $allergy(X, \text{"}Al\text{"}) \Rightarrow desease(X, \text{"}D\text{"})$
(support= α, confidence= β).

CONCLUSION

The paper discussed a method of generating multidimensional association rules. In general, multi-dimensional association rules consist of two types of rules, namely *interdimension association rules* and *hybrid-dimension association rules*. In this paper, we restricted our proposed method to generate inter-dimension association rules. Three pairs of equations were introduced to calculate support and confidence of three different kinds of generalized rules.

Advantage of the proposed method compared to the conventional one is that it can provide a more meaningful association rules by utilizing meaningful fuzzy labels. The problem of using fuzzy labels to provide the rules is generally how to assign appropriate membership degree for every fuzzy set.

In our next paper, we will discuss and propose a method to generate hybrid-dimension association rules by assuming that hybrid-dimension association rules is a hybridization between intradimension and interdimension association rules.

REFERENCES

1. Han J., Kamber, M., *Data Mining: Concepts and Techniques*, The Morgan Kaufmann Series, 2001.
2. Klir, G. J., Yuan, B., *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, New Jersey: Prentice Hall, 1995.
3. Intan, R., Mukaidono, M., "Toward a Fuzzy Thesaurus Based on Similarity in Fuzzy Covering", *Australian Journal of Intelligent Information Processing,* Vol.8 No.3, 2004. pp. 132-139.
4. Intan, R., Mukaidono M., "Generating Fuzzy Thesaurus by Degree of Similarity in Fuzzy Covering", *Proceedings of ISMIS 2003*, LNAI 2871, Springer-Verlag, 2003. pp. 427-432.
5. Intan, R., Mukaidono, M., "A Proposal of Fuzzy Thesaurus Generating by Fuzzy Covering", *Proceedings of NAFIPS 2003*, IEEE Press, 2003. pp. 167-172.

6. Gunawan, O. P., *Perancangan dan Pembuatan Aplikasi Data Mining dengan Konsep Fuzzy c-Covering untuk Membantu Analisis Market Basket pada Swalayan X*, (in Indonesian) Final Project, 2004.

7. Zadeh, L. A., "Fuzzy Sets and systems," *International Journal of General Systems*, Vol. 17, 1990. pp. 129-138.

8. Agrawal, R., Imielimski, T., Swami, A.N., "Mining Association Rules between Sets of Items in Large Database", *Proccedings of ACM SIGMOD International Conference Management of Data*, ACM Press, 1993. pp. 207-216.

9. Agrawal, R., Srikant, R., "Fast Algorithms for Mining Association Rules in Large Databases", *Proccedings of 20th International Conference Very Large Databases*, Morgan Kaufman, 1994. pp. 487-499.

10. Pesiwarissa, H. V., *Perancangan dan Pembuatan Aplikasi Data Mining dalam Menganalisa Track Records Penyakit Pasien di DR.Haulussy Ambon Menggunakan Fuzzy Association Rule Mining*, (in Indonesian) Final Project, 2005.

11. Intan, R., "An Algorithm for Generating Single Dimensional Fuzzy Association Rule Mining", *Jurnal Informatika,* Vol. 7, No. 1, Mei 2006.

12. Codd, E.F., "A Relational Model of Data for Large Shared Data Bank", *Communication of the ACM*, 13(6), 1970. pp. 377-387.

13. Intan, R., Mukaidono, M., 'Fuzzy Conditional Probability Relations and its Applications in Fuzzy Information System', *Knowledge and Information systems, an International Journal*, *Springer-Verlag*, Vol. 6(3), 2004.